the globus alliance
www.globus.org

GridFTP: Challenges in Bulk Data Movement

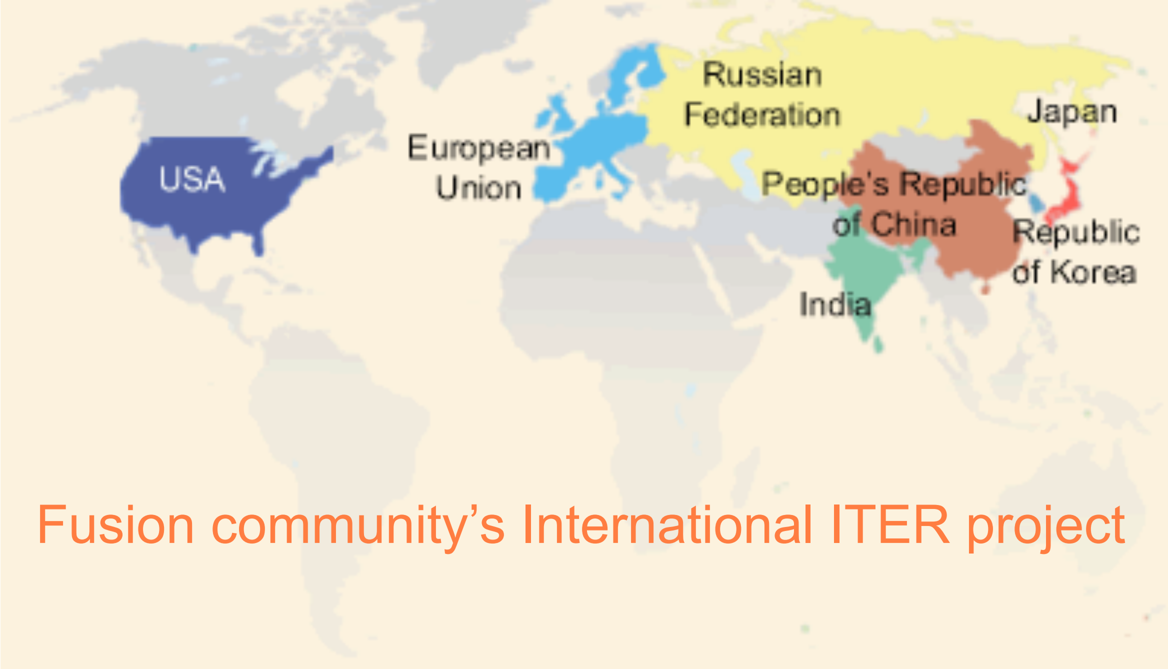
Raj Kettimuthu
Argonne National Laboratory and
The University of Chicago

Outline

- Introduction
- Network Capabilities
- End-to-End Problem
- Challenges
- GridFTP
- Future Directions

Today’s Science Environments

- Large-scale collaborative science is becoming increasingly common



Fusion community’s International ITER project

- Distributed community of users to access and analyze large amounts of data

Simulation Science

- In simulation science, the data sources are supercomputer simulations
  - For eg, DOE-funded climate modeling groups generate large reference simulations at supercomputer centers
- Combustion, fusion, computational chemistry, and astrophysics communities have similar requirements for remote and distributed data analysis

Experimental Science

- Data sources are facilities such as high energy and nuclear physics experiments and light sources.
  - For eg, LHC at CERN will produce petabytes of raw data per year for 15 years
- DOE light sources can also produce large quantities of data that must be distributed, analyzed, and visualized
- The international fusion experiment, ITER

Science Environments

- Raw simulation or observational data is just a starting point for most investigations
- Understanding comes from further analysis, reduction, visualization, and exploration

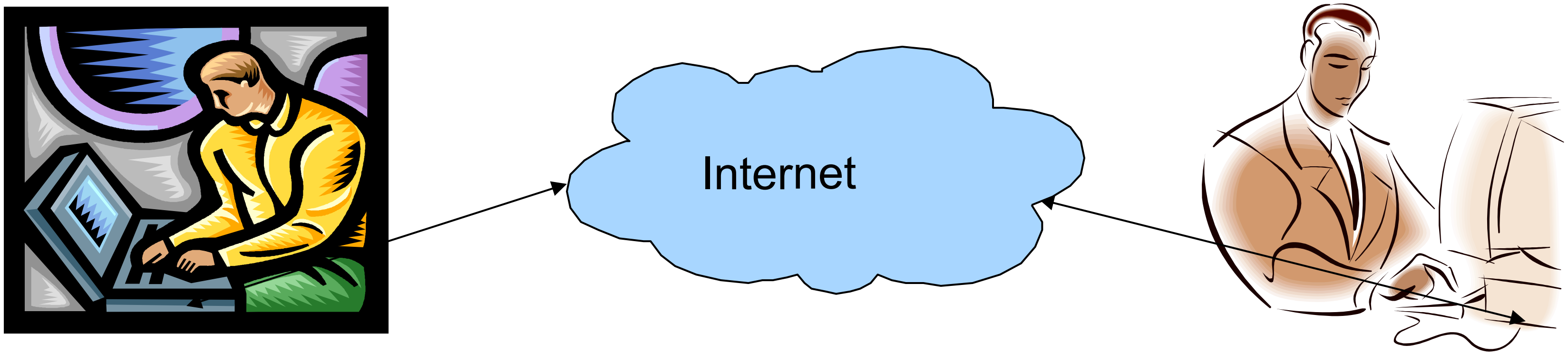
Petascale resource

Compute Cluster

Scientist’s Desktop

- Furthermore the data is a community asset that must be accessible to any member of a distributed collaboration

Network Capabilities



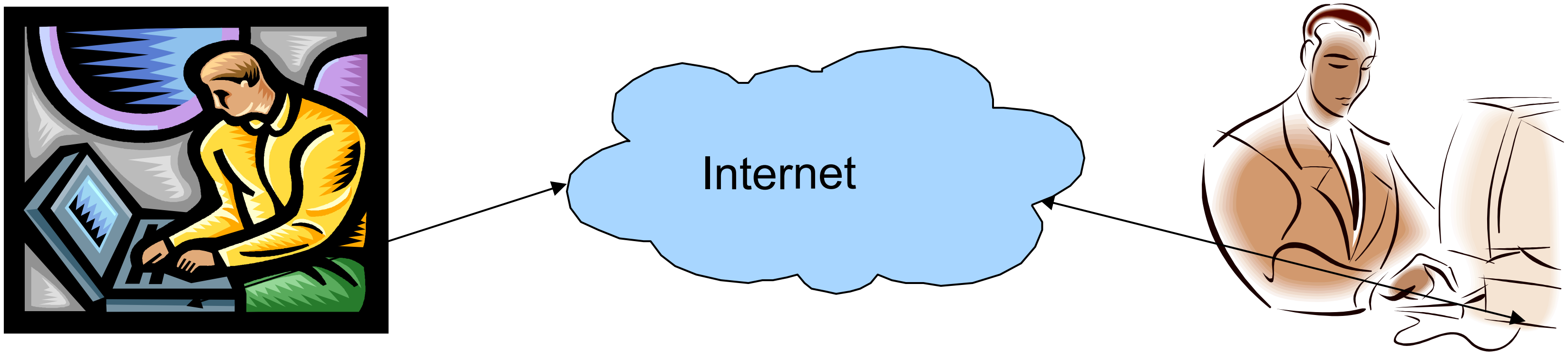
Internet

Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 1 Terabyte of data to Scientist B
- What is the fastest way to transfer the data?

Network Capabilities



Internet

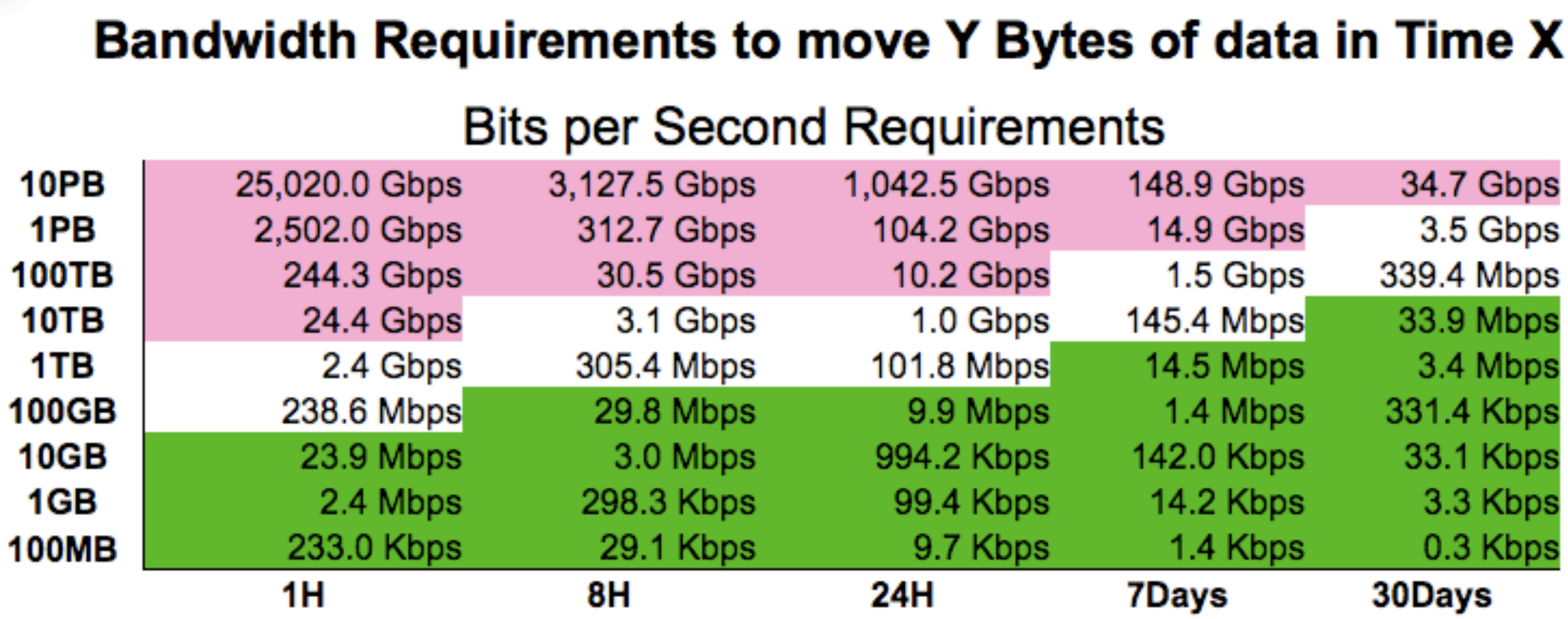
Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 1 Terabyte of data to Scientist B
- What is the fastest way to transfer the data?

FedEx

Bandwidth Requirements

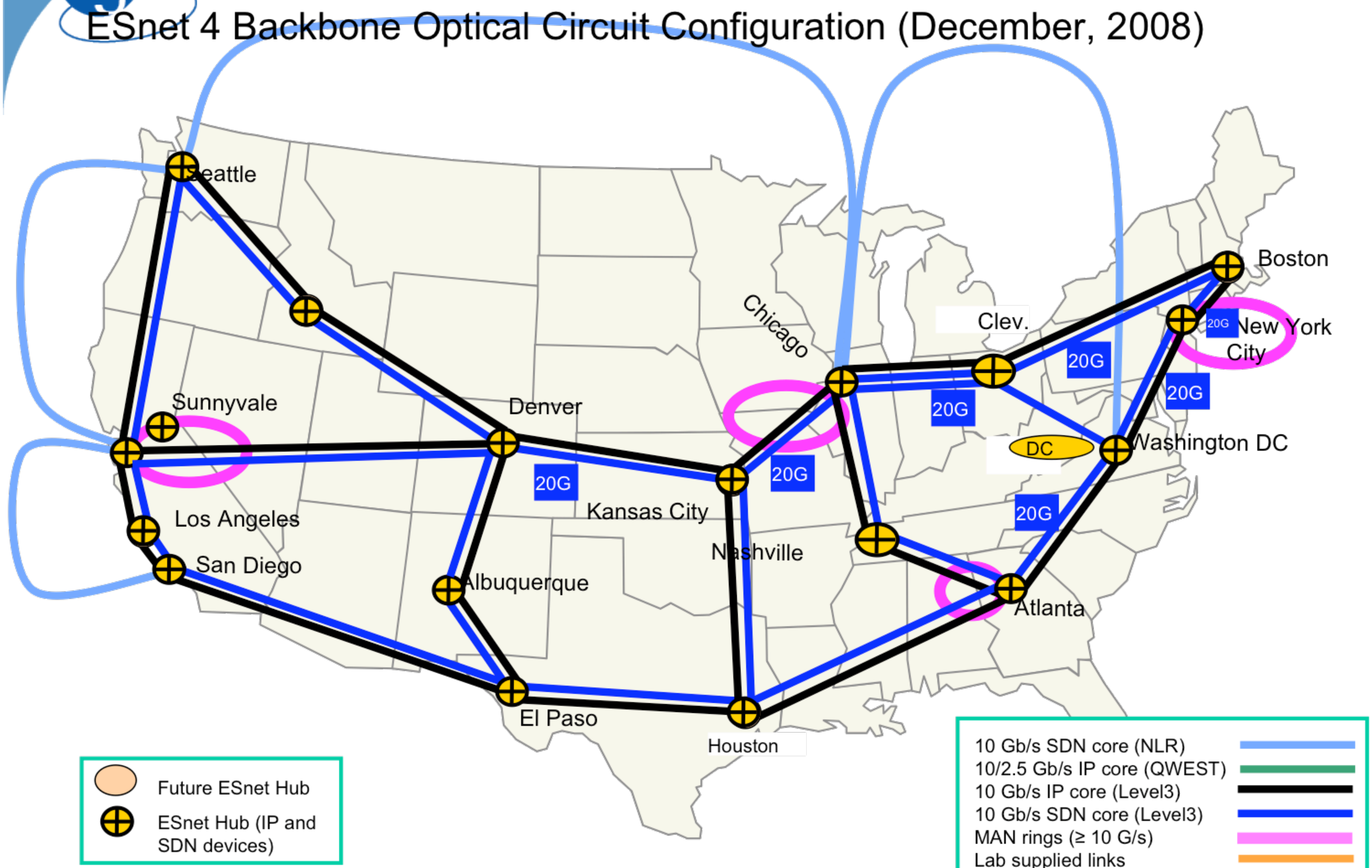


Bandwidth Requirements to move Y Bytes of data in Time X

Bits per Second Requirements

	1H	8H	24H	7Days	30Days
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps

ESNET



ESnet 4 Backbone Optical Circuit Configuration (December, 2008)

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross US in less than a second

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross US in less than a second

Yes. But it would still take two hours to get to downtown

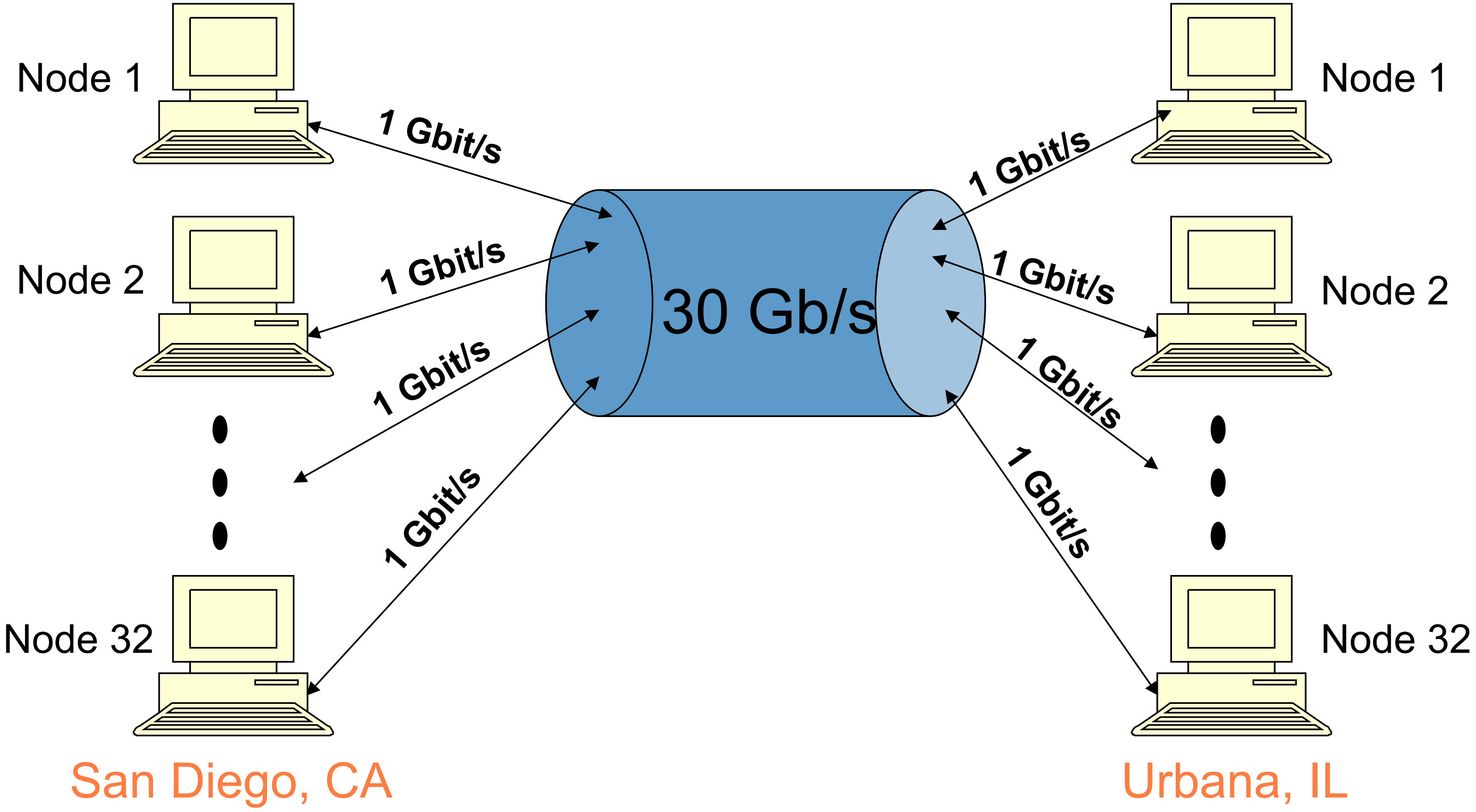
End-to-end problem

- Data movement in distributed science environments is an end-to-end problem
- A 10 Gbit/s network link between the source and destination does not guarantee an end-to-end data rate of 10 Gbit/s
- Other factors such as storage system, disk, data rate supported by the end node
- Deal with failures of various sorts
  - Firewalls can cause difficulties

the globus alliance
www.globus.org

End-to-end data transfer

Efficient and robust wide area data transport requires the management of complex systems at multiple levels.



Challenges

- Standard
- Throughput
- Robustness
- Secure
- Ease-of-use
- Scalable
- Extensible
- Reliable

GridFTP

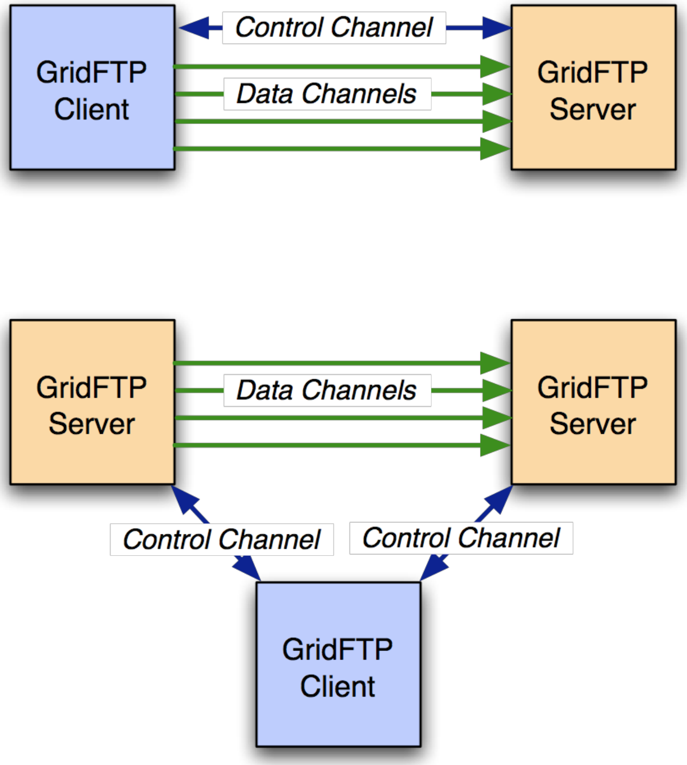
- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol - defines extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- GridFTP is the OGF recommended data movement protocol

GridFTP

- We (Globus Alliance) supply a reference implementation:
  - Server
  - Client tools
  - Development Libraries
- Multiple independent implementations can interoperate
  - Fermi Lab and U. Virginia have home grown servers that work with ours

GridFTP

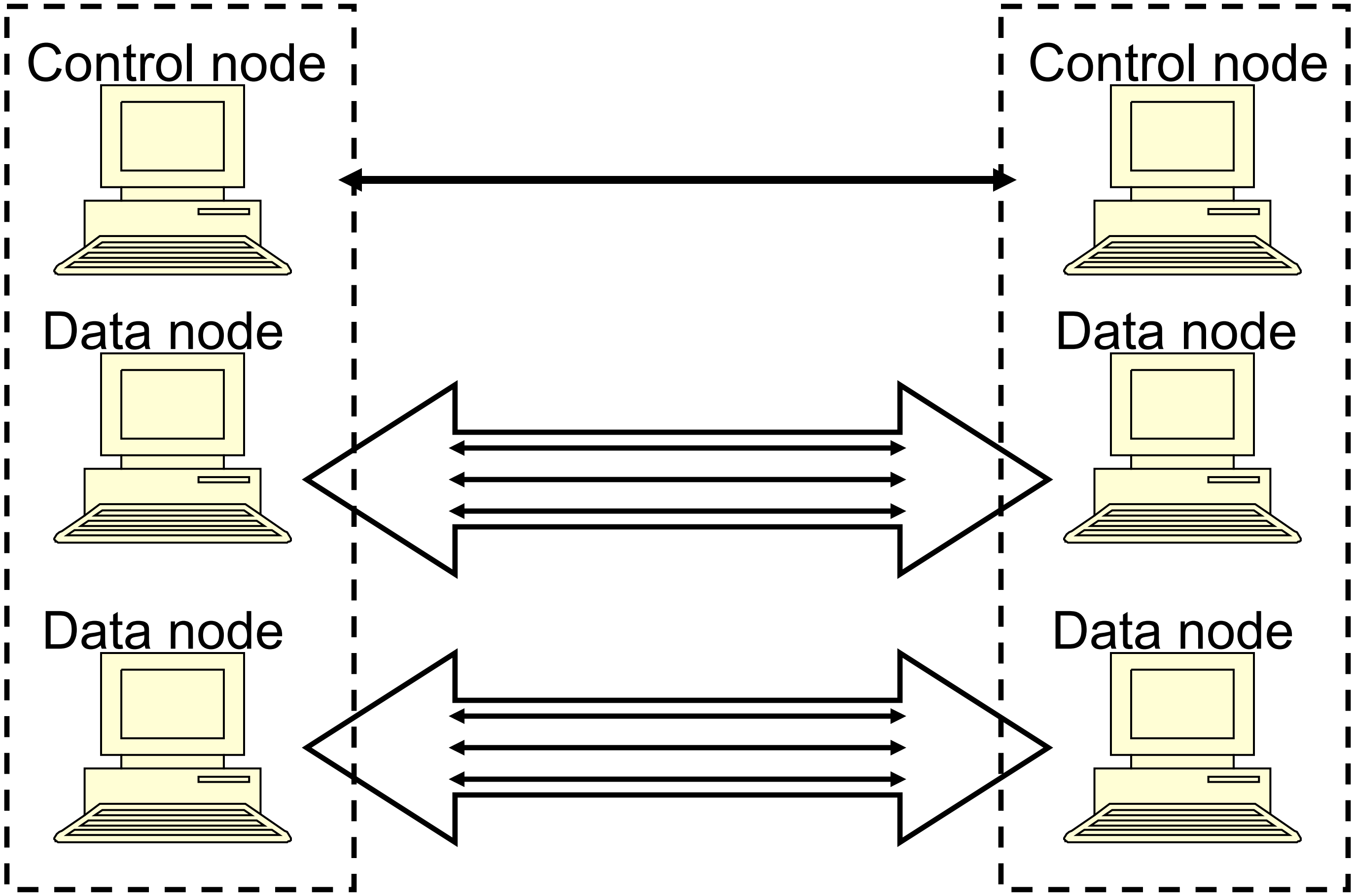
- Two channel protocol like FTP
- Control Channel
  - Communication link (TCP) over which commands and responses flow
  - Low bandwidth; encrypted and integrity protected by default
- Data Channel
  - Communication link(s) over which the actual data of interest flows
  - High Bandwidth; authenticated by default; encryption and integrity protection optional



Globus GridFTP Features

- GridFTP is Fast
  - Parallel TCP streams
  - Non TCP protocol such as UDT
  - Set optimal TCP buffer sizes
  - Order of magnitude greater
- Cluster-to-cluster data movement
  - Co-ordinated data movement using multiple computers at each end
  - Another order of magnitude

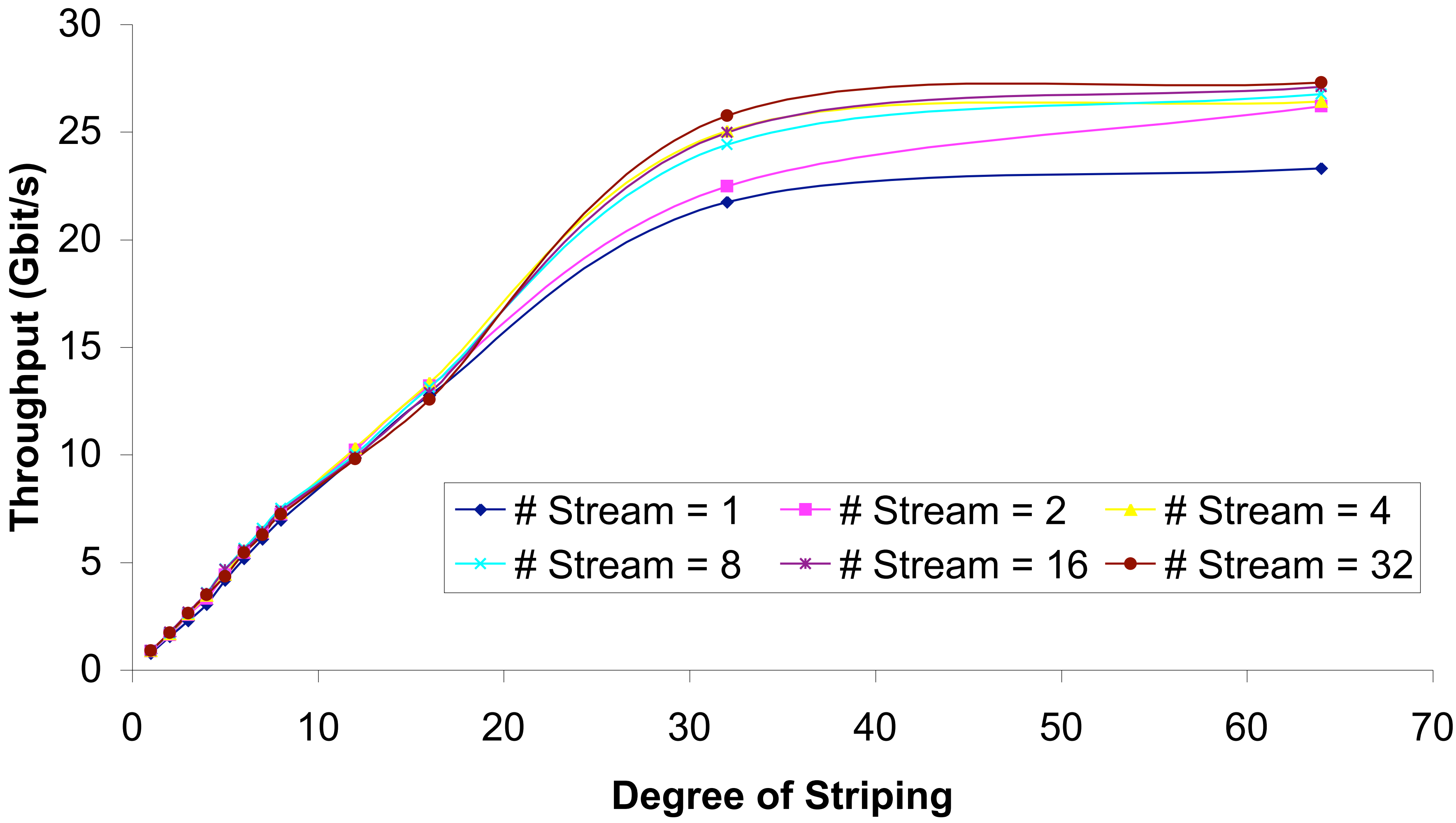
Cluster-to-Cluster transfers



the globus alliance
www.globus.org

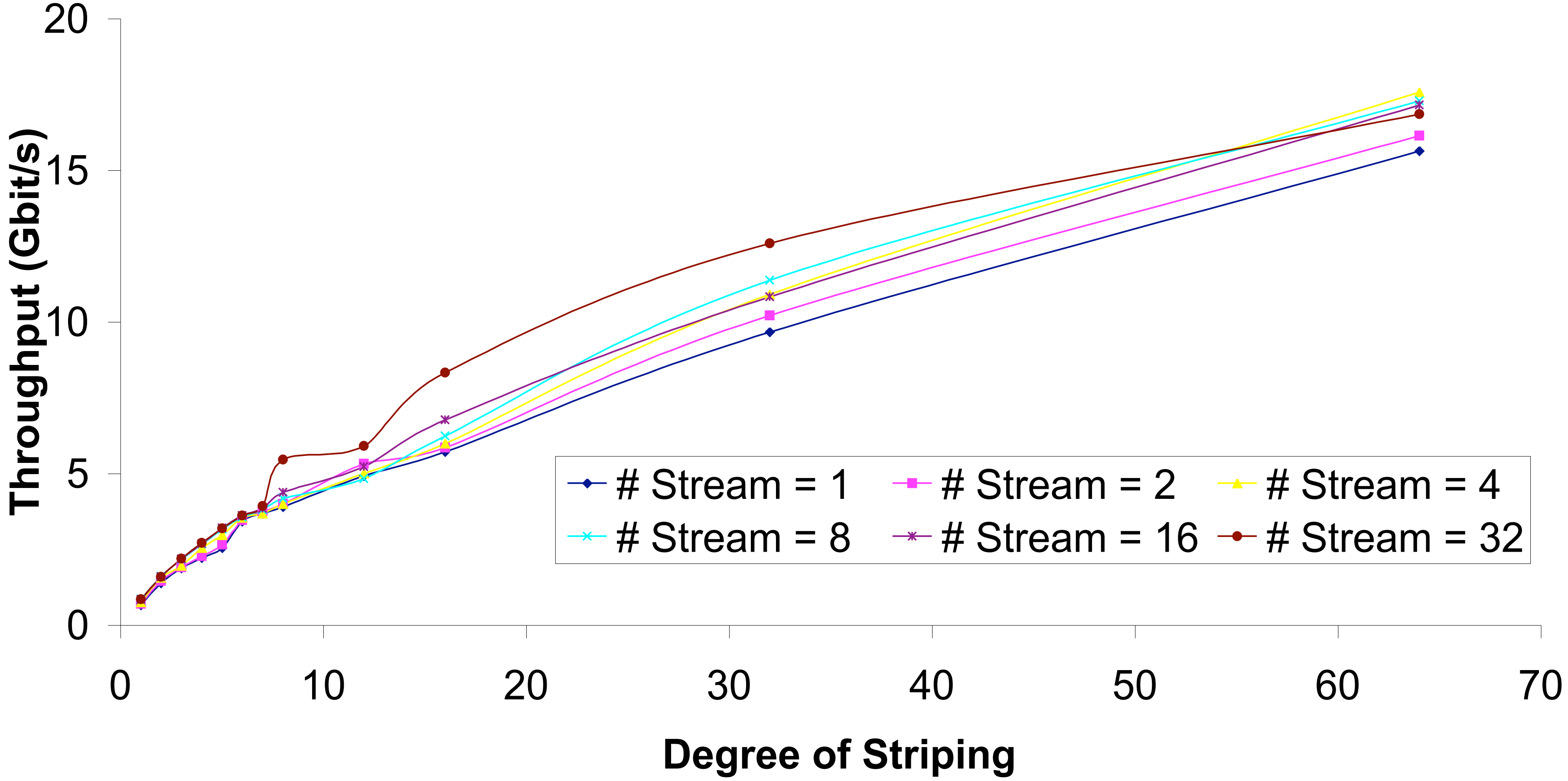
Performance

- Mem. transfer between Urbana, IL and San Diego, CA



Performance

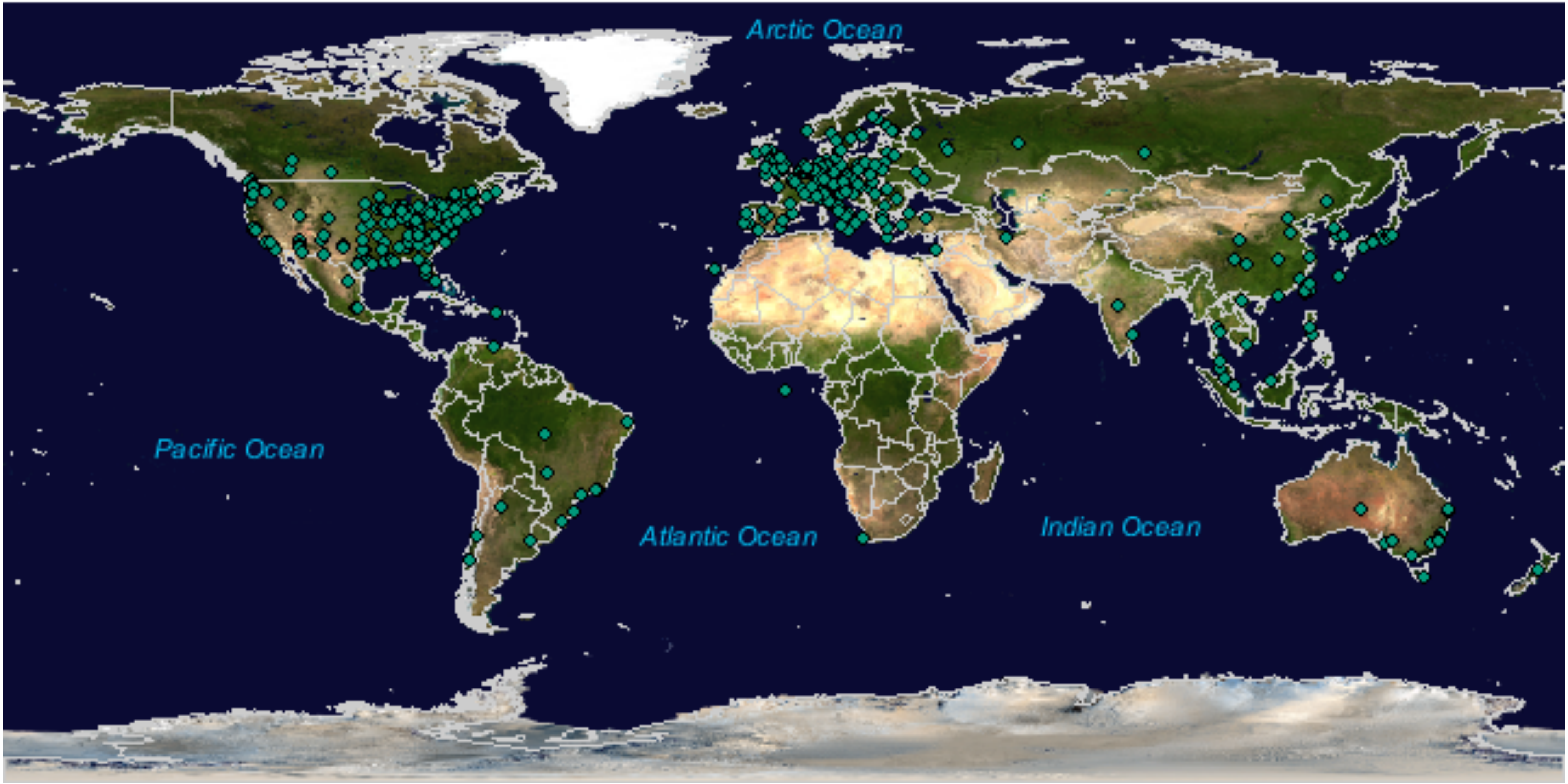
- Disk transfer between Urbana, IL and San Diego, CA



GridFTP in production

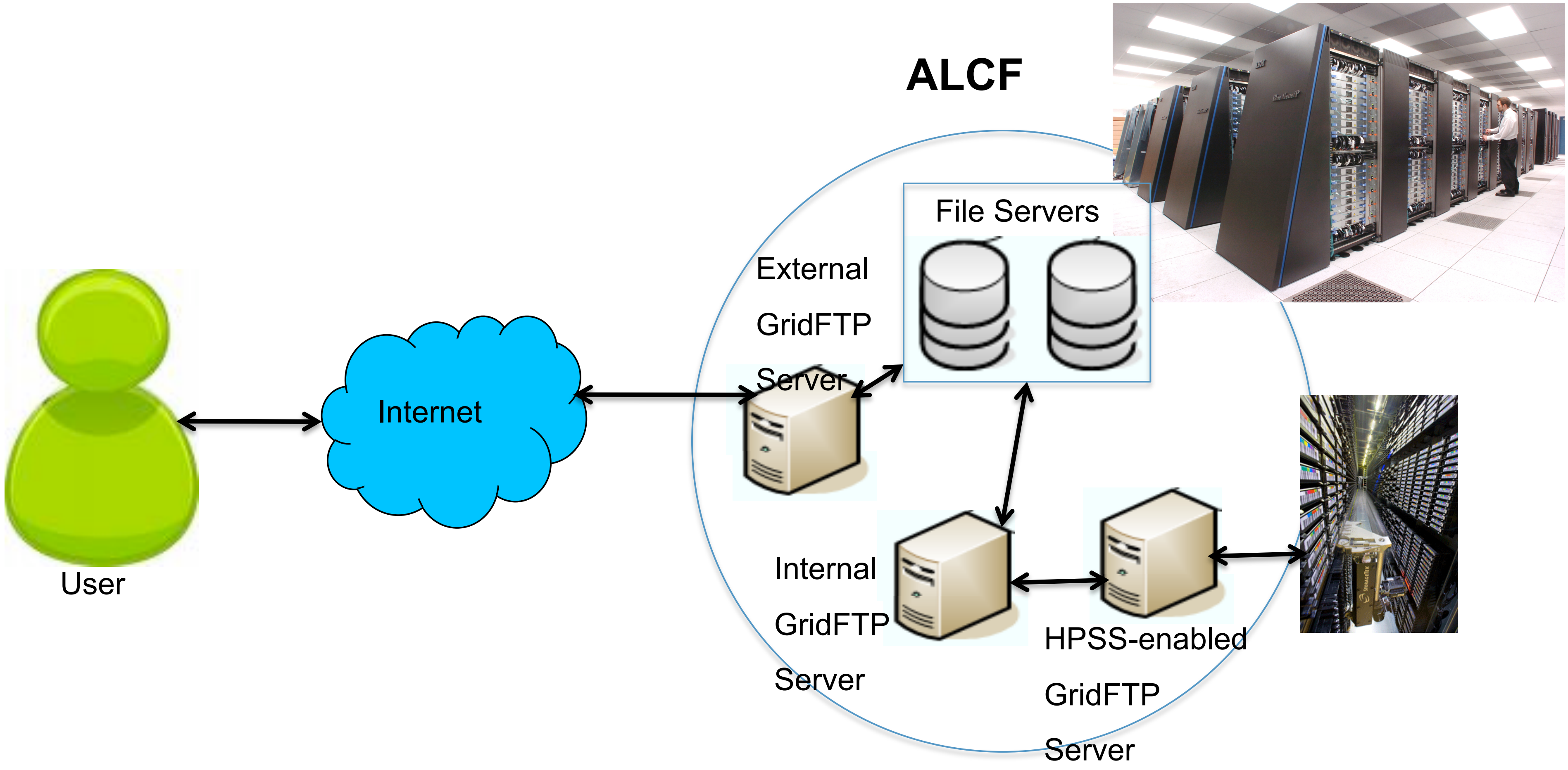
- Many Scientific communities rely on GridFTP
  - High Energy Physics - LHC computing Grid
  - Southern California Earthquake Center (SCEC), Earth Systems Grid (ESG), Relativistic Heavy Ion Collider (RHIC), European Space Agency, BBC use GridFTP for data movement
- GridFTP facilitates an average of more than 5 million data transfers every day

GridFTP Servers Around the World

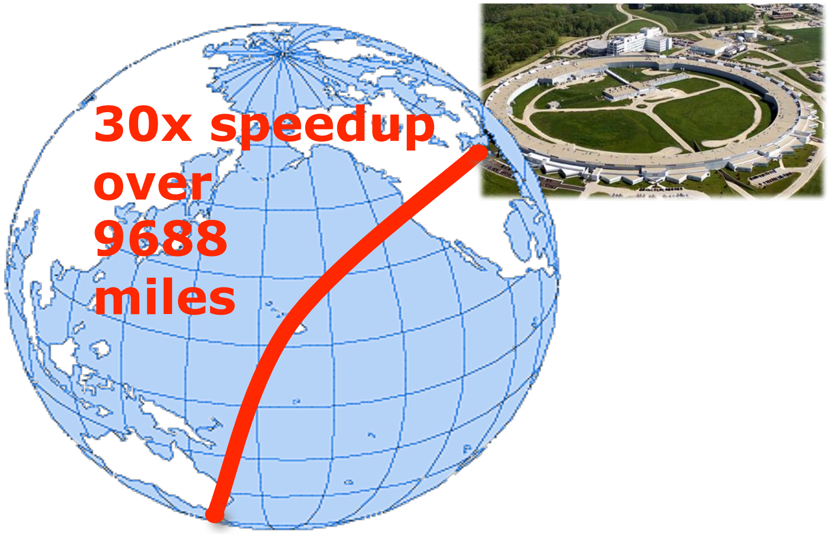


Created by Lydia Prieto ; G. Zarrate; Anda Imanitchi (Florida State University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

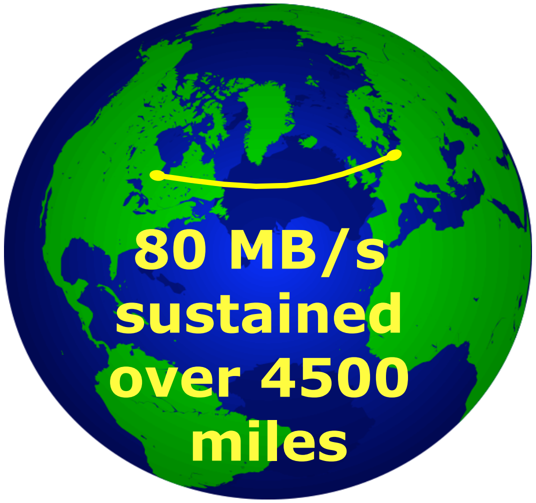
GridFTP in Production



GridFTP in production



One terabyte moved from an Advanced Photon Source tomography beamline to Australia, at a rate 30x faster than standard FTP

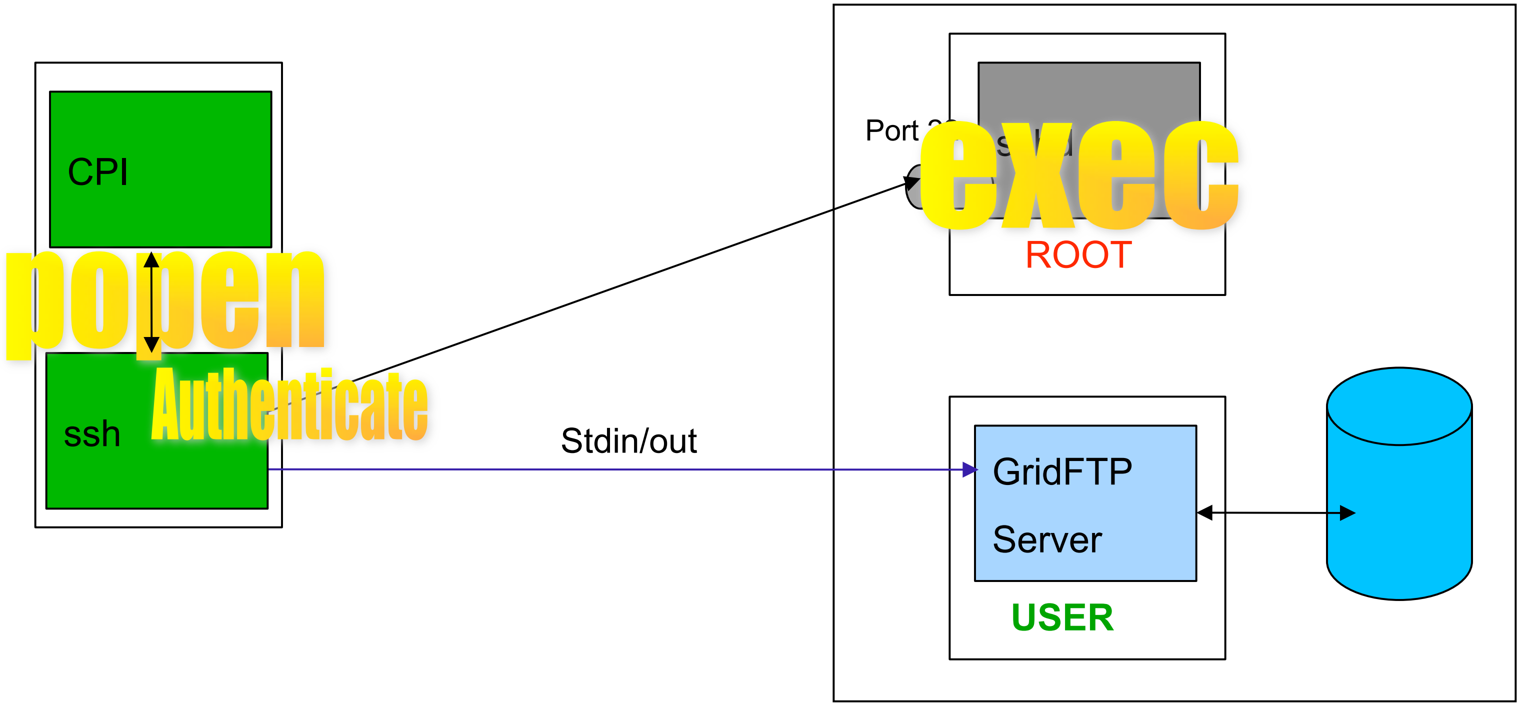


1.5 terabyte moved from University of Wisconsin, Milwaukee to Hannover, Germany at a sustained rate of 80 megabyte/sec

Security

- GridFTP provides strong security using GSI
- Protection vs. Ease of use
  - GSI and CAs were hard for many users
- Speed vs. protection
  - Users area happy with a minimal amount of data channel protection
- GridFTP over SSH
  - A big win for many users

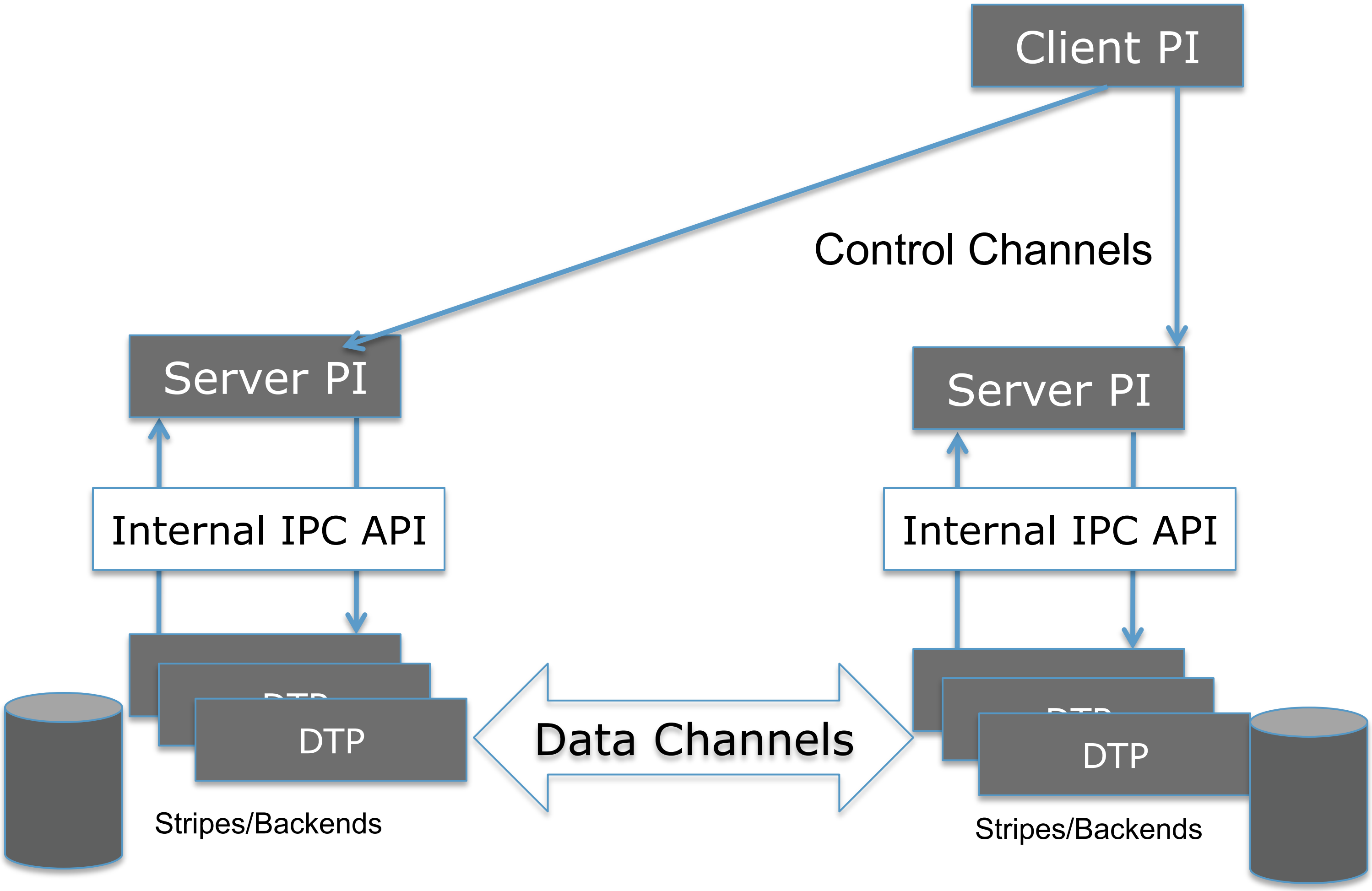
sshftp:// Interactions



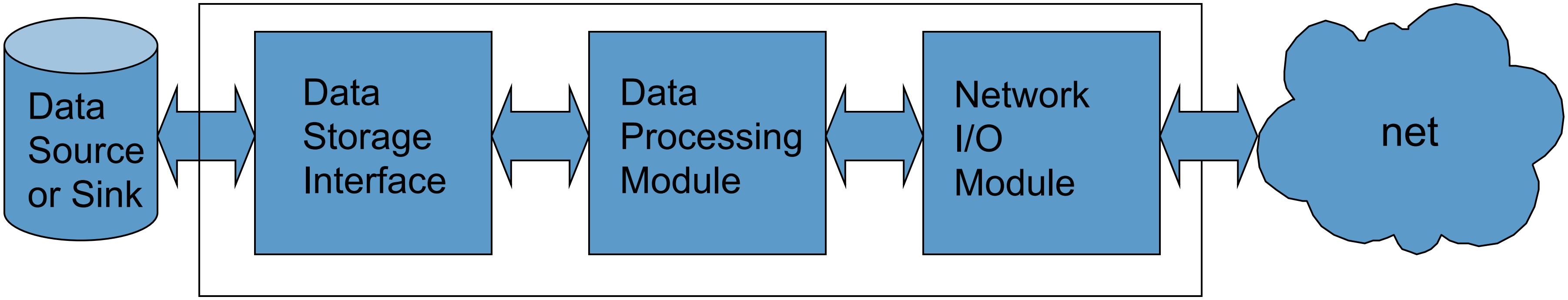
Challenges

- Past success
  - Standard – big selling point for adoption
  - Throughput – GridFTP was sold on speed
  - Robustness – has to work all the time
  - Secure – data channel security
- Current and future
  - Extensible
  - Reliable
  - Ease-of-use
    - Zero configuration clients
    - Firewall
  - Scalable

GridFTP Architecture



Modular



Well defined interfaces

Data Storage Interface (DSI)

- POSIX file system
- High Performance Storage System (HPSS)
- Storage Resource Broker (SRB)
- Hadoop DFS

Modular

- Network I/O module
  - Simple Open/Close/Read/Write interface
  - Well-defined abstraction called drivers
  - Easy to plug-in external libraries
  - TCP, UDT, Phoebus
- Data processing module
  - Compression (under development)
  - Checksum

Handling failures

- GridFTP server sends restart and performance markers periodically
  - Default every 5s - configurable
- Helpful if there is any failure
  - No need to transfer the entire file again
  - Use restart markers and transfer only the missing pieces
- GridFTP supports partial file transfers

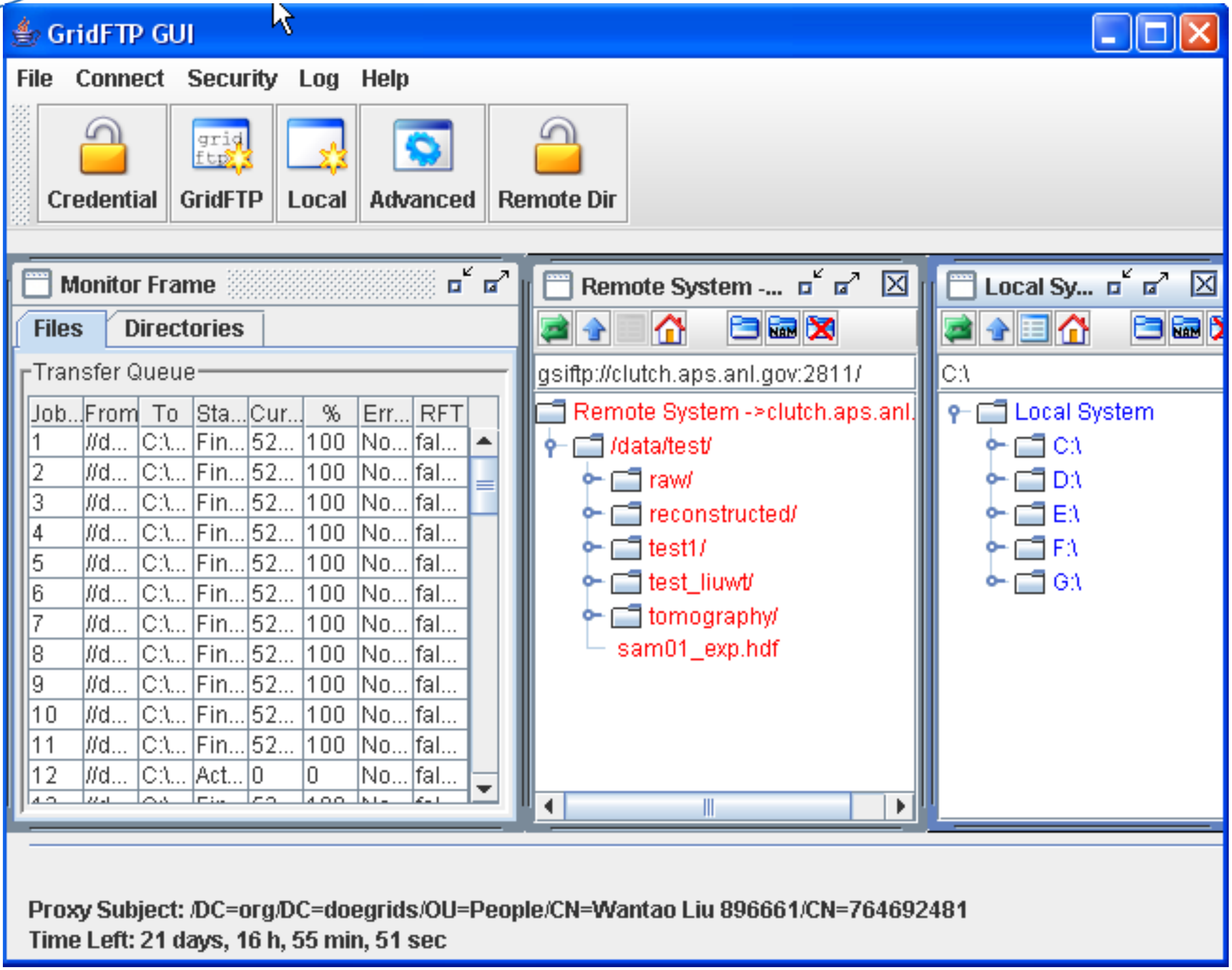
Handling failures

- Command-line client - globus-url-copy - support transfer retries
  - Use restart markers
- Recover from server and connection failures
- Improvements to globus-url-copy to recover from client failures

Easy-to-use

- Simple to install
    - Configure; make gridftp install;
    - Installs only gridftp and its dependencies
    - Binaries available for many platforms
- Various clients
    - Command-line client - globus-url-copy
    - Client libraries - well-defined API
    - Graphical User Interface

GridFTP GUI



Transfer Queue

Job...	From	To	Sta...	Cur...	%	Err...	RFT
1	//d...	C:\...	Fin...	52...	100	No...	fal...
2	//d...	C:\...	Fin...	52...	100	No...	fal...
3	//d...	C:\...	Fin...	52...	100	No...	fal...
4	//d...	C:\...	Fin...	52...	100	No...	fal...
5	//d...	C:\...	Fin...	52...	100	No...	fal...
6	//d...	C:\...	Fin...	52...	100	No...	fal...
7	//d...	C:\...	Fin...	52...	100	No...	fal...
8	//d...	C:\...	Fin...	52...	100	No...	fal...
9	//d...	C:\...	Fin...	52...	100	No...	fal...
10	//d...	C:\...	Fin...	52...	100	No...	fal...
11	//d...	C:\...	Fin...	52...	100	No...	fal...
12	//d...	C:\...	Act...	0	0	No...	fal...

gsiftp://clutch.aps.anl.gov:2811/

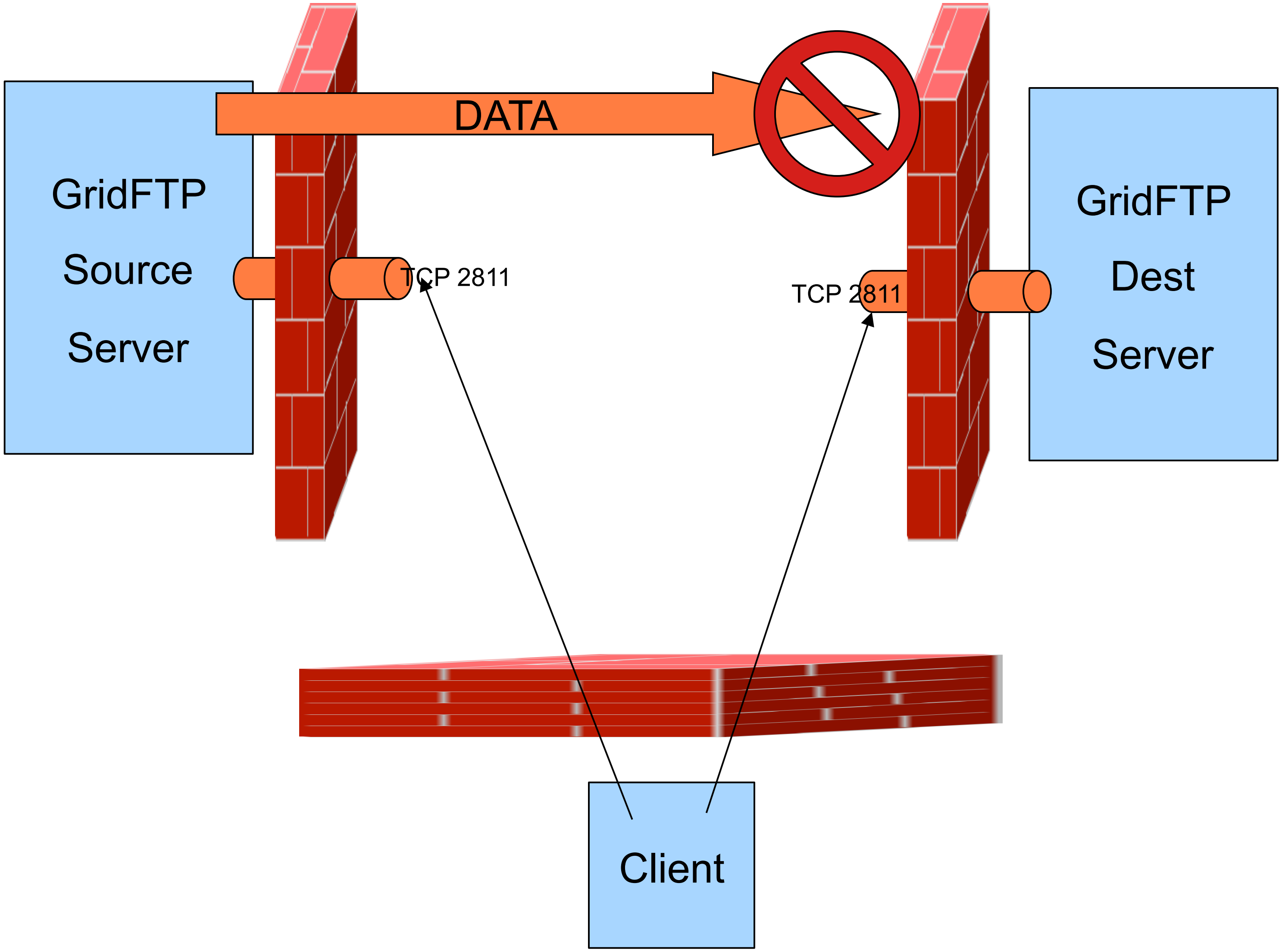
Remote System ->clutch.aps.anl.
/data/test/
raw/
reconstructed/
test1/
test_liuwt/
tomography/
sam01_exp.hdf

C:\

Local System
C:\
D:\
E:\
F:\
G:\

Proxy Subject: /DC=org/DC=doegrids/OU=People/CN=Wantao Liu 896661/CN=764692481
Time Left: 21 days, 16 h, 55 min, 51 sec

Firewalls


DATA
GridFTP
Source
Server
TCP 2811
TCP 2811
GridFTP
Dest
Server
Client

Firewalls

- Control channel is statically assigned
- Data channels dynamically assigned
- Single port GridFTP
  - Need to distinguish between the control channel and data channel
  - Need to associate data channels with the appropriate control channel
  - Backward compatibility is a challenge

Hosted Data Movement

- RFT evolution
  - Reliable File Transfer Service – WSRF based service
  - Configuration/setup not simple
- DataKoa
  - Hosted data movement service
  - Software as a service model
  - Fire and forget
    - Less user interaction
    - Email notifications

Questions